

DOCUMENT RESUME

ED 400 314

TM 025 689

AUTHOR Davey, Tim; And Others
TITLE Some New Methods for Mapping Ratings to the NAEP
Theta Scale To Support Estimation of NAEP Achievement
Level Boundaries.
PUB DATE Apr 96
NOTE 29p.; Paper presented at the Annual Meeting of the
National Council on Measurement in Education (New
York, NY, April 9-11, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) --
Speeches/Conference Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Achievement Tests; Elementary
Secondary Education; Error of Measurement;
*Estimation (Mathematics); Grade 4; Grade 12; Judges;
*Least Squares Statistics; Mathematics Achievement;
Maximum Likelihood Statistics; National Surveys;
*Performance; *Probability; Reading Achievement
IDENTIFIERS *Achievement Standards; Averaging (Mathematics);
Jackknifing Technique; Mapping; *National Assessment
of Educational Progress; Standard Setting

ABSTRACT

Some standard-setting methods require judges to estimate the probability that an examinee who just meets an achievement standard will answer each of a set of items correctly. These probability estimates are then used to infer the values on some latent scale that, in theory, determines an examinee's responses. The paper focuses on the procedures used to convert the probability estimates into performance standards. A number of procedures are described that have been traditionally used, including simple summation, simple ability averaging, weighted ability averaging, and least squares statistics. Some new procedures for estimating achievement standards from subjective probability estimates include an unweighted least squares approach, maximum likelihood (MLE), and modifications to MLE that consider posterior distributions. MLE, unweighted least squares, and simple ability averaging techniques were applied to data from the 1992 National Assessment of Educational Progress reading and mathematics tests for grades 4 and 12. These three approaches were evaluated with a jackknife design. Results for the different procedures were different, with lower achievement results and smaller standard errors from the MLE and least squares procedures than from the traditional simple ability averaging. The most desirable approach, however, may still depend on other than statistical criteria. (Contains one figure, six references, and five tables.) (SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

**Some New Methods
for Mapping Ratings to the NAEP Theta Scale
to Support Estimation
of NAEP Achievement Level Boundaries¹**

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as
received from the person or organization
originating it.
- ☐ Minor changes have been made to
improve reproduction quality.

- Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

Tim Davey
Meichu Fan
Mark Reckase
ACT

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY

Tim Davey

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

1. Introduction

It is often necessary to set minimum performance standards or passing scores for tests. These standards may be used to certify examinees, evaluate programs, or provide diagnoses (Shepard, 1984). Such standards are often determined judgmentally by asking groups of panelists or judges to review sets of test items (Angoff, 1971; Livingston & Zieky, 1982; Kane, 1987; Plake & Kane, 1991). Some standard setting methods require judges to estimate the probability that an examinee who just meets an achievement standard will answer each of a set of items correctly. These probability estimates are then used to infer the values on some latent scale that, in theory, determines an examinee's responses. The resulting values are typically aggregated across judges to set the minimum performance levels necessary for being considered as meeting an achievement standard.

Existing procedures differ more in the ways that judges actually estimate probabilities than in the ways these estimates are used in turn to set achievement standards. Under what is arguably the most commonly used procedure, judges are asked to envision the type and level of skills likely possessed by an examinee that just exceeds some subjective or criterion referenced standard (Angoff, 1971). For example, judges may be asked to consider examinees labeled as having "basic" or "proficient" skill levels. Since judges interpret these semantic labels in different ways, there is certain to be disagreement among judges from the outset. This is ordinarily resolved to some extent by training judges beforehand and encouraging discussion during the rating process. Agreement is also improved by having judges repeatedly revise their probabilities over several rounds with group discussion intervening.

The focus of this paper will not be on the process by which estimates of probabilities of correct response are produced, but rather on the somewhat more limited field of procedures used to convert probability estimates into performance standards. A number of such procedures are described below.

1.1 Simple summation.

Each judge's probability estimates, r_{ij} , are summed across items to produce the expected number-correct score that an examinee would have to equal or exceed to meet the achievement standard. These

¹Paper presented at the annual meeting of the National Council on Measurement in Education, New York, April, 1996.

scores are typically averaged across judges to yield an aggregate standard. Since the number-correct standard is relative to the rated pool of items, it must be adjusted to the extent that this pool differs from the actual test. This may be done using either item response theory or classical true score methods. With the former approach, number-correct scores are converted to values on the latent ability scale. Examinees whose ability estimates on the operational test exceed these values are deemed to have met the standard.

1.2 Simple ability averaging.

This procedure requires that the pool of rated items be fit by a common latent trait model. These models characterize item performance by *item response functions*, which give the (conditional) probability of correct response to each item by examinees with any fixed latent ability. The probability of a correct response to item i by an examinee with ability θ is denoted as $P_i(\theta)$. Since the judge's ratings, r_{ij} , are essentially estimates of these same probabilities conditional on examinees whose ability just meets the performance standard, the item response function can be used to convert each r_{ij} into a minimal latent ability value, $\hat{\theta}_{ij}$. For the ideal judge, each item probability would convert to the same latent ability. In practice, judges are not nearly so consistent and item response functions are estimated rather than known. The values of the inferred latent abilities therefore often vary considerably across items even with experienced judges. The judges themselves also generally differ considerably from one another in the standards that they apply and the consistency with which they apply them. Accordingly, inferred latent ability values are usually averaged across items and judges to produce an aggregate standard, θ_0 . Alternatively, the passing probabilities can be averaged across judges for each item and these averages transformed to latent abilities through test characteristic curves, which are themselves averaged item characteristic curves, to produce the aggregate standard (Kane, 1987).

Methods based on simple sums or averages of the item probabilities allow all judges and items to contribute equally toward determining the aggregate standard. However, since judges vary both in their degree of internal consistency, and to the extent that they agree with other judges, and items differ in how well their responses functions are estimated, treating all item and judges as equally valid may be a questionable practice.

1.3 Weighted ability averaging.

These methods recognize that the quality of information provided differs across judges and items by producing aggregate standards as weighted rather than simple averages. Various weighting functions have been proposed, with the general effect being to down weight the contributions of internally inconsistent judges and/or noninformative or poorly modeled items (Kane, 1987).

1.4 Least squares.

Kane (1987) proposed setting the aggregate standard as the latent ability value that yields a "best fit" to the judges reported passing probabilities on each item. More formally, θ_0 is found that minimizes:

$$\sum_{j=1}^N \sum_{i=1}^n (r_{ij} - P_i(\theta_0))^2 \quad (1)$$

A weighted least squares version of the procedure was also proposed in which the squared difference terms

are weighted by the reciprocals of the variances of the probabilities for each item across judges:

$$w_i = \frac{1}{\sigma_i^2(r_{ij})} \quad (2)$$

Under both types of least squares procedures, items and judges contribute unequally to determination of the aggregate standard. However, differences in contributions are more pronounced with the weighted version, which explicitly down weights items with ratings that show considerable variability across judges.

Plake and Kane (1991) compared the simple ability averaging, weighted ability averaging, and least squares methods with a simulation study. Results were somewhat equivocal, which led to a recommendation for the more parsimonious simple ability averaging method.

2. Some new methods

Several new procedures for estimating achievement standards from subjective probability estimates have been developed to better meet the needs of achievement level setting for the National Assessment of Educational Progress (NAEP). To improve the behavior of these estimates, the ratings are first transformed to the logit metric:

$$r'_{ij} = \ln \frac{r_{ij}}{1 - r_{ij}} \quad (3)$$

The logit transformation was chosen empirically after applying a variety of transformations to actual data. It was found to yield the most nearly normal distribution of r'_{ij} , and also to equalize the variances across

judges of $r'_{ij} - \text{logit } P_i(\theta_0)$, the errors in prediction of the observed item probabilities from the estimated aggregate standard. Both of these results are important to the effectiveness or efficiency of the new estimation procedures.

2.1 Unweighted least squares.

This first procedure is similar to that proposed by Kane (1987), with the important difference that logit-transformed item probability estimate rather than the observed estimates are used. The logit transformation is crucial because unweighted least squares estimation directly assumes equality of the error variances across items and will lose efficiency to the extent that these variances differ. The objective function minimized by the unweighted least squares procedure is:

$$\min_{\theta_0} \sum_{j=1}^N \sum_{i=1}^n (r'_{ij} - \text{logit } P_i(\theta_0))^2 \quad (4)$$

This function can be minimized by iterative numerical methods.

2.2 Maximum likelihood.

This procedure is based on the assumption that each judge's transformed probability ratings $\langle r'_1, r'_2, \dots, r'_n \rangle$ constitute a sample from a joint distribution, f^θ , parameterized by an ability value θ . The objective is then to find the ability parameter that maximizes the likelihood of the observed ratings having occurred. More concretely, let:

$$f_i^\theta = N(\mu_i(\theta), \sigma_i^2) \quad (5)$$

where

$$\mu_i(\theta) = \text{logit } P_i(\theta) = \ln \frac{P_i(\theta)}{1 - P_i(\theta)} \quad (6)$$

$$\sigma_i^2 = \text{Var}(r'_i - \mu_i(\theta)) \quad (7)$$

The aggregate standard is then estimated by finding the ability value that maximizes the likelihood of the observed ratings over items and judges:

$$L(\theta | r'_j) = \prod_{i=1}^n \prod_{j=1}^N f_i^\theta(r'_j) \quad (8)$$

The MLE procedure relaxes the assumption of equal variances by allowing σ_i^2 to differ across items.

However, since it was found impractical to estimate variances separately for each item, items are instead sorted into four homogeneous subsets and a common variance estimated for the item group. The four subsets consist of items assumed to have small, moderately small, moderately large, and large variances, respectively. The following procedure is used to assign items to variance groups:

1. Map each observed probability estimate to an ability value through the item characteristic curve. This results in a distribution of ability values for each judge. The spread of this distribution is an indicator of a judge's consistency across items. A consistent judge will predict response probabilities that each map to nearly the same ability value. This value is the location on the latent scale of the examinee or group of examinees that define the achievement level.
2. Estimate the latent achievement value for each rater, $\hat{\theta}_j$, as the median of the distribution of inferred ability values obtained in step 1, above.
3. Find the logit transformation of each rating, r'_{ij} . Also take the logit transformation of the rating "predicted" for each item and examinee by the item response model parameters:

$$\hat{r}'_{ij} = \text{logit } P_i(\hat{\theta}_j, a_i, b_i, c_i) \quad (9)$$

4. Compute the "residual" variance for each item as:

$$\hat{\sigma}_i^2 = \sum_{j=1}^N (r'_{ij} - \hat{r}_{ij})^2 \quad (10)$$

5. Sort items into four groups on the basis of their residual variances.

Iterative numerical methods are used to find θ_0 and four values of σ^2 (one for each item group) that maximize the likelihood (8). Maximization can be done for either individual examinees or jointly for multiple examinees.

2.3 Posterior distributions.

The maximum likelihood procedure can be modified and theoretically improved in two fundamental ways:

1. **Produce posterior distributions of achievement standards rather than point estimates.** The ML procedure yields a point estimate of the "true" achievement standard. This estimate summarizes (by its mode) some distribution of plausible achievement standards. Unfortunately, the quality of this summary may differ widely from application to application. For example, the mode may not adequately characterize the location of a highly skewed distribution. The alternative is then to produce the entire distribution and then determine the statistic that provides the best summary.

The posterior distribution is defined at each ability value as:

$$h(\theta) = \frac{f^{\theta}(r_1, r_2, \dots, r_n) g(\theta)}{\int f^{\theta}(r_1, \dots, r_n) g(\theta) d\theta} \quad (11)$$

where $g(\theta)$ is some prior distribution on the "true" achievement levels. While the same prior distribution could be used regardless of the achievement level being estimated, a better approach may be to adjust the prior depending on achievement level being estimated.

2. **Relax the implicit assumption that each judge rates items with respect to the same true achievement standard.** The ML procedure estimates a single achievement standard jointly across all raters. While the procedure can estimate a unique level for each individual rater, it is not clear how these estimates are best combined to yield a single level. An alternative is to instead sum the individual posterior distributions across raters to construct a single, joint posterior distribution. There are at least two options to do so:
 - a. Weight judges equally when forming the joint posterior. Under this option, the individual posteriors are simply summed over examinees.

- b. Weight judges unequally when forming the joint posterior, thereby acknowledging that some judges ratings are more coherent than others. For example, posteriors can be weighted inversely to their variance when the summed or joint distribution is being produced.

3. Extensions to polytomous items

Polytomous or partial credit items are generally scored on a multi-point scale. Judges can be asked to rate these items by stating the expected score for an examinee at a given achievement level. Estimation of achievement standards from polytomous items can then be handled by procedures identical to those outlined above, with two important differences:

1. Judged item scores are transformed to the (0-1) scale by dividing each by the maximum scale score. The logit transformation is then applied to the result to produce r'_{ij} .
2. Multiple category items are calibrated under a polytomous latent trait model such as the generalized partial credit model (Muraki, 1992). Under this model, expected item scores $P_i(\theta_j)$ are given by:

$$P_i(\theta_j) = \sum_{m=1}^M m \frac{\exp \left[\sum_{k=1}^m a_i(\theta_j - b_i + d_{ik}) \right]}{\sum_{n=1}^M \exp \left[\sum_{l=1}^n a_i(\theta_j - b_i + d_{il}) \right]} \quad (12)$$

The logit transformation is again applied to yield the r'_{ij} .

4. Evaluation

Both the maximum likelihood and unweighted least square procedures were applied to data from the 1992 NAEP Reading and Mathematics Grade 4 and 12, Group A standard setting studies. For comparison purposes, achievement levels were also estimated using the simple ability averaging procedure recommended by Plake and Kane (1991). Three achievement levels (basic, proficient, and advanced) were estimated for Mathematics and Reading for each of two grades (fourth and twelfth), resulting in a total of twelve estimation conditions. For convenience, analysis was restricted to dichotomous items only. Twelve judges rated 90 (grade 4) and 105 (grade 12) mathematics items, respectively. Eleven judges rated the 47 (grade 4) and 59 (grade 12) reading items.

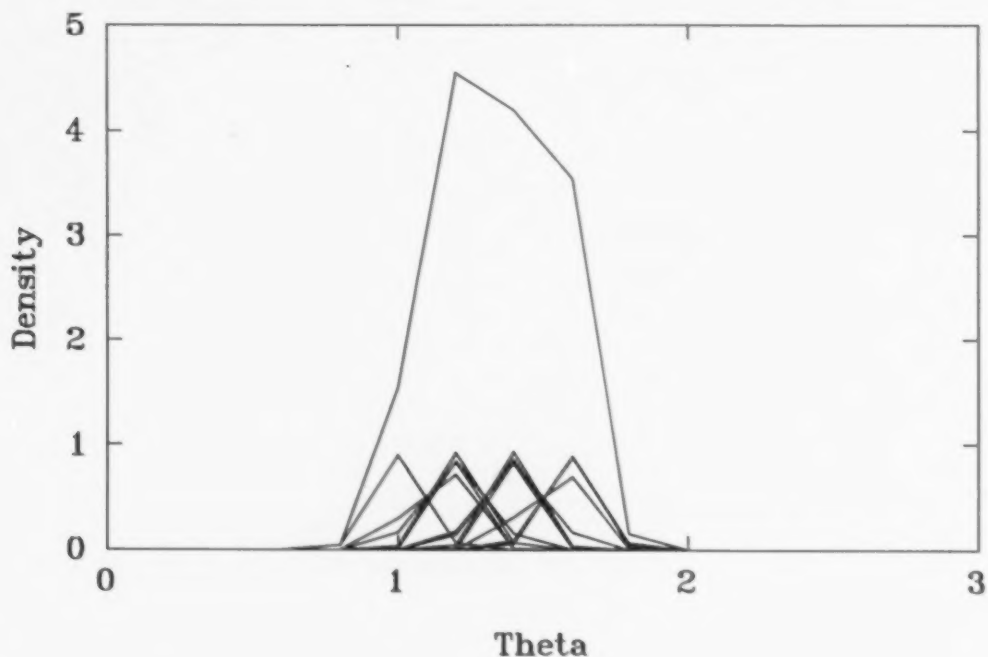
The three estimation procedures were evaluated with a jackknife design under which individual judges were successively dropped from the data set. The result of this design is a set of n estimates of each standard, each based on $n-1$ judges. The stability of an estimation procedure is indicated by corresponding stability of the set of estimates. A stable procedure will be relatively unaffected by the small perturbations of the data sample induced by the removal of a single judge. In this case, the set of estimates should have relatively small variance. In contrast, an unstable procedure will yield very different, and consequently quite variable, estimates following minor changes in the data sample.

4.1 Results

Table 1 shows the achievement level estimates and jackknife variances for the Reading and Mathematics data sets. There is a minor but fairly consistent tendency for the simple ability averaging procedure to produce higher standard estimates than either the LS or MLE procedures. The MLE estimates tend to be smallest, with LS somewhere in between. Much more apparent is the greater stability demonstrated by the LS and MLE estimates when compared to simple ability averaging. The LS jackknife variance was smallest (most stable) under eight of the twelve conditions. The MLE variances were smallest in three conditions, with ability averaging appearing most stable only once. The smallest variance under each condition is highlighted in the tables.

In addition to the common achievement levels estimated jointly from groups of judges, levels were also estimated separately for individual judges. Such estimates would be used as a form of feedback to judges during the successive stages of the rating process. The individual estimates for the Grade 4 Mathematics study are shown in Table 2. Corresponding results for the Grade 4 Reading study are in Table 3. Only minor differences in the rank orderings of examinees across estimation procedures were found.

The equally weighted variant of the posterior distribution procedure was also applied as an example to data from the 1994 Geography, Grade 4, Group B study. The Advanced achievement level was estimated from both dichotomous and polytomous item simultaneously. The results are shown on the graph below. The smaller distributions at the bottom of the plot are from each of the fourteen judges present in the study. The larger distribution is the equally-weighted aggregate distribution.



The maximum likelihood procedure furnished an estimate of 1.379 from these same data, a value that lies squarely in the center of the aggregate distribution.

4.2 Item misfit statistics

Judges typically determine achievement standards through a multistage process that involves repeatedly estimating the probabilities of items being answered correctly. Following each round of estimates, judges are often presented with feedback that shows each how they stand with respect to other judges. Judges are also shown a list of items that are "misfit" in the sense that the probabilities estimated for these items differ considerably from the probabilities expected given the achievement standard inferred from the entire set of items. Items can be misfit by their probabilities being either overestimated or underestimated. Table 4 shows each judge's five most overestimated items for grade 4 Mathematics study. Table 5 shows the underestimated items from this same data set. The general conclusion drawn from these tables is that the MLE and LS procedures yield results that are far more similar to one another than either is to unweighted ability averaging procedure.

5. Discussion

The selection of a procedure for mapping judges ratings of items to an IRT scale is surprisingly difficult when the possible procedures are thoroughly analyzed. The traditional method for performing this mapping is to average the judges ratings and map them to the IRT scale using the test characteristic curve. This is the method called Simple Ability Averaging in this paper. However, this process, though having the appeal of simplicity, has a number of technical problems. First, the test characteristic curve is regression of test score on θ . This is the IRT analog to the linear regression of Y on X . Continuing with the analogy to linear regression, the goal of standard setting is to predict X from Y . This is best done with a different regression function than Y on X , which is $X = f(Y)$, that minimizes the error in estimating X . The function that is needed for standard setting is the regression of θ on X , rather than the function we have readily available, the test characteristic curve.

Mapping backward through the test characteristic curve may not lead to much increase in error over the use of opposite regression if the test characteristic curve is well estimated. If the correlation between the variables in linear regression is close to 1.0, the two different regression functions are very similar. However, if the test characteristic curves are poorly estimated, the increase in error from using the wrong function could be substantial.

A second technical issue is whether the same IRT model is being used to map the ratings that is being used to estimate the examinees' proficiency levels on the test. When item ratings are summed to produce a score, the implicit IRT model is a Rasch model since the total score is a sufficient statistic for θ under that model. If the model used to estimate examinees' proficiency is not the Rasch model, as is the case in NAEP, there is a mismatch in the models. The results of the mappings will be different for the two models to the extent that the estimated item characteristic curves differ across the models.

The ideal mapping procedure is one that would use a procedure that is consistent with the estimation procedure for examinees' proficiency, one that yields results with small errors of estimation, and one that has a strong theoretical rationale. In this paper, several procedures have been presented that attempt to meet these criteria. The Maximum Likelihood and Posterior Distribution procedures provide a closer match to the proficiency estimation procedure used in NAEP than do the other procedures. They also seek to minimize the error in estimated θ rather than to use the test characteristic curve. Finally, the

procedures use the three-parameter logistic model rather than an implied Rasch model.

The results for the different procedures are, not surprisingly, different. The Maximum Likelihood and Least Squares procedures tend to give lower achievement level estimates and smaller standard errors than the traditional procedure. This is probably due to the fact that these procedures do not weight all observations equally. Erratic judges and poorly discriminating items tend to be weighted less, stabilizing the estimates. Whether the differential weighting is desirable depends on other than statistical criteria. It may be unacceptable to weight the ratings of one judge less than another and differential weighting of items may have subtle effects on the weighting of content in the achievement levels. These issues will need to be settled in a non-technical forum. The hope for this paper is that it will at least make the issues clear so that they can be given active consideration.

6. References

- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.
- Livingston, S. A. & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.
- Kane, M. T. (1987). On the use of IRT models with judgmental standard setting procedures. *Journal of Educational Measurement*, 24(4), 333-345.
- Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159-176.
- Plake, B. S. & Kane, M. T. (1991). Comparison of methods for combining the minimum passing levels for individual items into a passing score for a test. *Journal of Educational Measurement*, 28(3), 249-256.
- Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.) *A guide to criterion-referenced test construction*. Baltimore: The Johns Hopkins University Press.

Table 1

Achievement level estimates and jackknife variances for Reading, Grade 4

Level		MLE	LS	Averaging
Basic	Standard	-1.315	-1.172	-1.303
	Variance	.019	.015	.023
Proficient	Standard	-.500	-.470	-.393
	Variance	.015	.013	.016
Advanced	Standard	.201	.193	.267
	Variance	.021	.023	.024

Achievement level estimates and jackknife variances for Reading, Grade 12

Level		MLE	LS	Averaging
Basic	Standard	-.303	-.188	-.367
	Variance	.021	.016	.027
Proficient	Standard	.624	.627	.686
	Variance	.006	.006	.008
Advanced	Standard	1.408	1.414	1.651
	Variance	.028	.031	.031

Achievement level estimates and jackknife variances for Mathematics, Grade 4

Level		MLE	LS	Averaging
Basic	Standard	-.127	.011	-.182
	Variance	.042	.035	.051
Proficient	Standard	.833	.832	.971
	Variance	.035	.032	.043
Advanced	Standard	1.719	1.708	2.025
	Variance	.051	.063	.042

Table 1 (cont.)

Achievement level estimates and jackknife variances for Mathematics, Grade 12

Level		MLE	LS	Averaging
Basic	Standard	-.130	-.145	-.050
	Variance	.026	.022	.038
Proficient	Standard	.823	.910	1.030
	Variance	.019	.017	.025
Advanced	Standard	1.684	1.691	1.873
	Variance	.025	.025	.041

Table 2

Basic Achievement Level Estimates for Individual Judges, Mathematics, Grade 4

Judge	MLE	LS	Averaging
1	.456 (1)	.559 (1)	.548 (1)
2	.097 (5)	.231 (5)	.095 (5)
3	.218 (3)	.346 (3)	.198 (3)
4	.382 (2)	.489 (2)	.391 (2)
5	.177 (4)	.330 (4)	.182 (4)
6	.013 (6)	.182 (6)	-.064 (6)
7	-.474 (9)	-.273 (9)	-.658 (9)
8	-.032 (7)	.154 (7)	-.089 (7)
9	-.474 (9)	-.291 (10)	-.658 (9)
10	-.812 (11)	-.693 (11)	-1.297 (11)
11	-1.027 (12)	-.814 (12)	-1.555 (12)
12	-.295 (8)	-.172 (8)	-.503 (8)

Table 2 (Continued)

Proficient Achievement Level Estimates for Individual Judges, Mathematics, Grade 4

Raters	MLE	LS	Averaging
1	1.131 (2)	1.266 (2)	1.643 (1)
2	1.337 (1)	1.238 (1)	1.617 (2)
3	1.190 (4)	1.162 (4)	1.443 (4)
4	1.192 (3)	1.169 (3)	1.445 (3)
5	.756 (8)	.800 (7)	.859 (9)
6	.798 (6)	.802 (6)	.974 (6)
7	.597 (10)	.618 (10)	.660 (10)
8	.887 (5)	.890 (5)	1.02 (5)
9	.749 (9)	.762 (9)	.867 (8)
10	.416 (11)	.416 (11)	.425 (11)
11	.072 (12)	.109 (12)	.071 (12)
12	.781 (7)	.755 (8)	.906 (7)

Table 2 (Continued)

Advanced Achievement Level Estimates for Individual Judges, Mathematics, Grade 4

Raters	MLE	LS	Averaging
1	2.165 (4)	2.137 (4)	2.461 (4)
2	2.513 (1)	2.513 (1)	2.673 (1)
3	2.475 (2)	2.484 (2)	2.672 (2)
4	2.197 (3)	2.187 (3)	2.532 (3)
5	1.140 (11)	1.287 (10)	1.652 (10)
6	1.537 (8)	1.536 (8)	1.952 (8)
7	1.615 (7)	1.614 (7)	1.989 (7)
8	1.999 (5)	2.001 (5)	2.239 (5)
9	1.486 (9)	1.461 (9)	1.827 (9)
10	.992 (12)	.977 (12)	1.228 (12)
11	1.241 (10)	1.215 (11)	1.526 (11)
12	1.973 (6)	1.955 (6)	2.226 (6)

Table 3

Basic Achievement Level Estimates for Individual Judges, Reading, Grade 4

Raters	MLE	LS	Averaging
1	-1.399 (7)	-1.251 (7)	-1.421 (6)
2	-1.144 (3)	-.992 (3)	-1.037 (3)
3	-1.501 (11)	-1.399 (11)	-1.647 (11)
4	-1.370 (6)	-1.221 (6)	-1.431 (7)
5	-1.186 (4)	-1.106 (4)	-1.211 (4)
6	-1.290 (5)	-1.130 (5)	-1.284 (5)
7	-1.478 (10)	-1.239 (8)	-1.436 (8)
8	-1.468 (9)	-1.327 (10)	-1.441 (9)
9	-1.466 (8)	-1.303 (9)	-1.571 (10)
10	-1.038 (1)	-.953 (1)	-.999 (1)
11	-1.134 (2)	-.978 (2)	-1.025 (2)

Proficient Achievement Level Estimates for Individual Judges, Reading, Grade 4

Raters	MLE	LS	Averaging
1	-.581 (8)	-.537 (7)	-.473 (7)
2	-.204 (1)	-.141 (1)	-.142 (1)
3	-.363 (2)	-.347 (3)	-.182 (2)
4	-.650 (10)	-.635 (11)	-.610 (11)
5	-.482 (6)	-.478 (6)	-.400 (6)
6	-.574 (7)	-.557 (8)	-.500 (10)
7	-.735 (11)	-.622 (10)	-.545 (8)
8	-.609 (9)	-.593 (9)	-.571 (9)
9	-.476 (5)	-.447 (5)	-.351 (5)
10	-.438 (4)	-.397 (4)	-.305 (4)
11	-.376 (3)	-.306 (2)	-.217 (3)

Table 3 (cont.)

Advanced Achievement Level Estimates for Individual Judges, Reading, Grade 4

Raters	MLE	LS	Averaging
1	.208 (6)	.234 (5)	.319 (6)
2	.567 (1)	.632 (1)	.667 (1)
3	.023 (9)	-.067 (10)	.070 (9)
4	.163 (7)	.160 (6)	.210 (7)
5	.210 (5)	.151 (7)	.296 (5)
6	.059 (8)	.039 (8)	.105 (8)
7	-.140 (11)	-.118 (11)	-.045 (11)
8	-.031 (10)	-.018 (9)	.021 (10)
9	.391 (3)	.403 (2)	.525 (3)
10	.274 (4)	.295 (4)	.416 (4)
11	.432 (2)	.402 (3)	.585 (2)

Table 4

Overestimated Items for Each Judge, Grade 4 Mathematics Study
Advanced Achievement Level

Judge	MLE	LS	Averaging
1	15 72 7 55 83	15 72 89 3 29	19 72 2 15 7
2	72 82 63 7 83	72 2 82 19 81	19 2 72 11 81
3	72 62 7 63 83	72 62 89 7 63	2 72 7 19 8
4	72 7 15 55 81	72 2 81 7 11	2 19 72 11 7
5	62 72 7 82 32	62 72 7 82 32	7 72 11 62 2
6	72 55 28 61 83	72 28 55 61 83	19 72 84 28 15
7	55 59 43 70 61	55 59 43 2 32	2 8 5 72 13
8	72 83 82 3 43	72 3 83 82 43	19 72 3 7 2

9	72	72	72
	55	55	19
	61	38	2
	38	61	84
	32	32	38
10	82	82	19
	61	61	2
	15	15	82
	70	32	15
	83	70	5
11	70	70	72
	72	72	19
	85	85	8
	63	63	63
	74	32	85
12	43	43	19
	72	29	2
	22	72	11
	15	23	72
	63	19	13

Table 4 (cont.)

Overestimated Items for Each Judge, Grade 4 Mathematics Study
Proficient Achievement Level

Judge	MLE	LS	Averaging
1	72	72	72
	55	55	15
	27	27	7
	15	61	9
	61	15	19
2	61	61	72
	28	28	11
	72	55	2
	55	59	28
	59	72	7
3	61	61	2
	72	72	72
	55	75	5
	59	55	81
	27	59	7
4	72	72	72
	61	61	2
	55	81	81
	81	55	7
	59	59	5
5	72	72	72
	32	32	5
	43	43	80
	80	80	7
	59	59	82
6	61	61	72
	55	55	28
	72	72	7
	59	32	61
	32	59	55
7	59	59	2
	55	55	9
	61	61	72
	32	9	8
	9	32	13
8	59	59	72
	72	32	59
	32	72	9
	61	61	70
	67	43	43

9	32 55 61 72 27	32 55 61 72 27	72 19 2 32 55
10	61 32 59 67 32	61 32 9 77 75	9 28 70 32 61
11	54 77 32 53 72	32 77 29 54 9	9 29 70 72 4
12	61 32 72 55 67	32 61 72 55 69	72 2 9 32 28

Table 4 (cont.)

Overestimated Items for Each Judge, Grade 4 Mathematics Study
Basic Achievement Level

Judge	MLE	LS	Averaging
1	77 61 55 72 32	61 77 55 72 32	72 45 77 55 79
2	61 55 59 79 72	61 55 59 75 43	72 70 28 55 29
3	61 55 32 59 72	61 55 32 59 72	72 61 55 32 29
4	61 55 72 32 59	61 55 72 32 75	72 61 55 75 59
5	72 43 61 32 55	72 43 61 32 55	72 43 80 37 70
6	61 55 32 72 59	61 55 72 32 75	72 61 55 70 69
7	59 53 61 54 32	59 61 53 54 32	59 69 61 67 32
8	54 32 59 53 61	61 43 54 59 32	72 70 59 32 43

9	53 61 54 32 59	61 53 59 32 54	59 61 32 75 16
10	68 54 32 53 61	68 32 54 61 53	32 68 67 75 23
11	54 68 53 61 57	68 54 61 53 57	68 90 35 59 54
12	32 54 53 68 44	32 54 43 53 44	32 43 69 38 72

Table 5

Underestimated Items for Each Judge, Grade 4 Mathematics Study
Advanced Achievement Level

Judge	MLE	LS	Averaging
1	54	54	47
	74	74	46
	47	47	33
	46	46	20
	86	86	17
2	53	53	47
	58	58	46
	74	74	17
	65	42	58
	35	35	20
3	65	68	58
	68	42	46
	53	53	53
	42	65	47
	54	54	65
4	57	57	20
	65	65	10
	68	68	12
	12	25	33
	58	10	17
5	53	53	47
	35	30	80
	30	35	85
	85	85	10
	32	54	33
6	53	53	46
	57	57	12
	46	46	47
	54	54	40
	12	12	16
7	58	46	46
	46	58	12
	12	53	7
	53	54	47
	74	57	17
8	53	53	73
	54	54	17
	35	35	20
	65	73	24
	58	65	58

9	47	47	21
	57	57	47
	58	20	20
	21	58	8
	35	46	46
10	58	58	17
	74	53	58
	47	54	73
	53	57	45
	65	74	47
11	58	58	33
	57	57	21
	53	33	58
	55	53	17
	33	54	57
12	57	57	58
	58	58	47
	86	86	46
	53	54	17
	54	53	10

Table 5 (cont.)

Underestimated Items for Each Judge, Grade 4 Mathematics Study
Proficient Achievement Level

Judge	MLE	LS	Averaging
1	46	46	17
	47	47	47
	20	55	46
	58	61	20
	86	15	21
2	58	58	20
	53	53	17
	20	20	10
	46	46	12
	47	47	47
3	46	46	11
	47	47	47
	58	58	17
	53	20	46
	86	86	20
4	58	58	19
	57	12	10
	12	10	12
	10	20	17
	46	57	20
5	17	17	85
	58	85	11
	20	20	17
	85	58	13
	46	73	20
6	46	46	2
	57	20	11
	20	12	10
	47	10	16
	58	57	20
7	46	46	19
	12	12	28
	47	42	7
	58	7	31
	37	24	12
8	58	17	17
	46	58	20
	88	20	10
	86	46	73
	17	24	24

9	46	21	21
	47	20	20
	21	46	17
	58	47	15
	20	58	10
10	47	17	19
	58	47	17
	46	12	8
	17	58	47
	86	46	12
11	88	20	61
	80	44	87
	47	33	28
	55	10	38
	58	88	27
12	47	47	10
	58	58	47
	46	10	17
	10	20	20
	53	12	12

Table 5 (cont.)

Underestimated Items for Each Judge, Grade 4 Mathematics Study
Basic Achievement Level

Judge	MLE	LS	Averaging
1	46	46	28
	47	47	19
	33	33	13
	86	20	17
	58	17	66
2	20	20	19
	47	17	2
	58	10	7
	44	47	13
	86	46	17
3	47	47	19
	86	17	28
	58	46	11
	74	12	17
	76	20	31
4	58	12	19
	47	10	28
	57	85	11
	86	17	85
	76	58	8
5	58	85	89
	76	17	87
	87	20	18
	47	33	85
	17	13	27
6	20	20	19
	47	46	2
	58	10	11
	76	17	8
	86	12	13
7	81	81	31
	44	73	2
	28	44	66
	73	28	19
	52	17	50
8	17	17	19
	86	12	2
	58	20	66
	47	10	55
	81	9	28

9	86 87 52 66 79	86 52 87 79 84	2 21 79 67 36
10	47 81 52 87 79	26 88 12 17 20	11 2 34 31 50
11	72 44 87 47 69	17 44 12 88 72	5 11 84 32 2
12	81 47 86 58 79	81 47 49 37 20	19 9 18 86 36